Visual Object Recognition

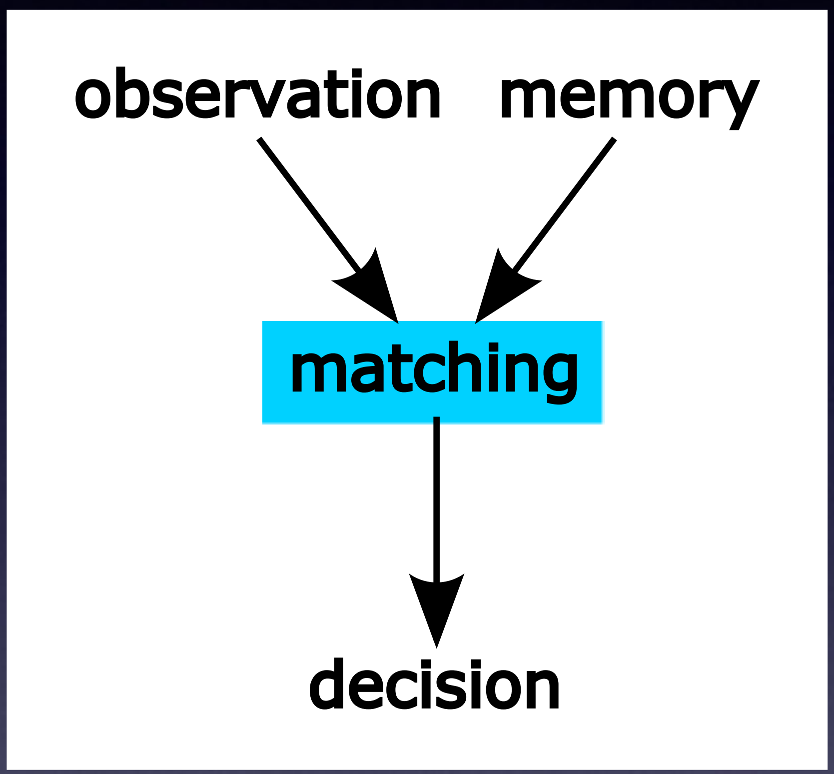
Lecture 5: Compound Descriptors and Metrics

Per-Erik Forssén, docent
Computer Vision Laboratory
Department of Electrical Engineering
Linköping University

Seminar 8 date

- All seminars shifted by one week.
- Exception:
  LE8 will take place on Wednesday March 25
  12.30-15.

Lecture 5: Compound Descriptors and Metrics



- Until now we have focused on how to construct the observation.
- This lecture is about how to arrange observations for matching.
- We will also look at similarity, and distance measures.

Lecture 5: Compound Descriptors and Metrics

- Feature Constellations
- Bags of Features and Visual Words
  Feature Sampling, Spatial Pyramids
- Descriptor distances
  Chi^2 distance, Earth Mover's Distance (EMD)
- Ratio Score Matching
- Learning the metric

Feature constellations

- Both **Local appearance** and **constellations** contribute to the recognition process.
- Case study of visual agnosia:
  Oliver Sacks, ”The man who misstook his wife for a hat”, 1985

Feature constellations

The Librarian

Vertumnus, Rudolf II

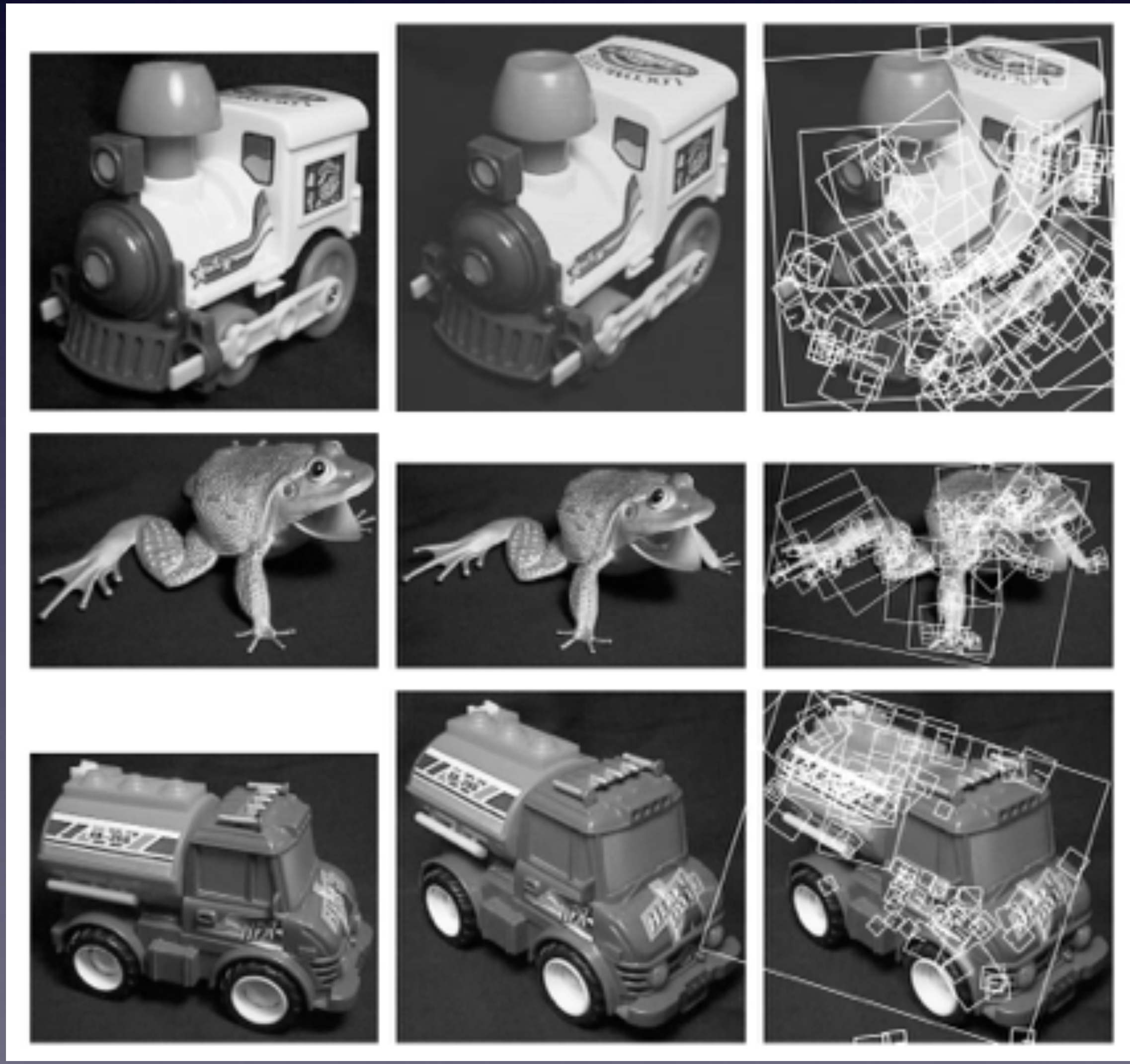
- Italian painter Giuseppe Arcimboldo 1527-1593 exploited how constellations inform recognition

Feature constellations

- D.G. Lowe, ”Local Feature View Clustering for 3D Object Recognition”, CVPR’01

- A ***view based*** object representation.

- An object is a set of views. In each view an affine transform constrains the feature constellation.



Feature constellations

- D.G. Lowe, ”Local Feature View Clustering for 3D Object Recognition”, CVPR’01
- During learning, similar views are clustered into fewer, if they can agree on a feature arrangement under an affine transformation.
- As 3D geometry is not explicitly used, views can represent both pose changes and articulation of the object.

Feature constellations

- D.G. Lowe, ”Local Feature View Clustering for 3D Object Recognition”, CVPR’01
- In recognition, matching is first made by having each feature in the query image vote for matching views.
- Views are then verified using the affine constellation model.
- Scales to many objects using ANN-trees (LE6), but eventually trees become too large.

Bags of features

- Another order of magnitude can be handled by Bags of features (introduced in todays paper)
  J. Sivic and A. Zisserman, ”Video Google: A text retrieval approach to object matching in videos”, ICCV’03

Illustration by Li Fei-Fei, http://people.csail.mit.edu/torralba/shortCourseRLOC/

Visual Words

- Closely related to Bags of Keypoints, Bags of features (BoF), Bags of words (BoW), and Texton histograms.
  G. Csurka et al, ”Visual Categorization with Bags of Keypoints”, ECCV’04
- Used for quickly indexing large datasets.
- Completely disregards spatial relationships among features.
- Spatial arrangement should be verified in a second step.

Visual Words

- Descriptor space (e.g. SIFT) is vector quantized into K parts on large training set.

- Clustering is done in whitened space:

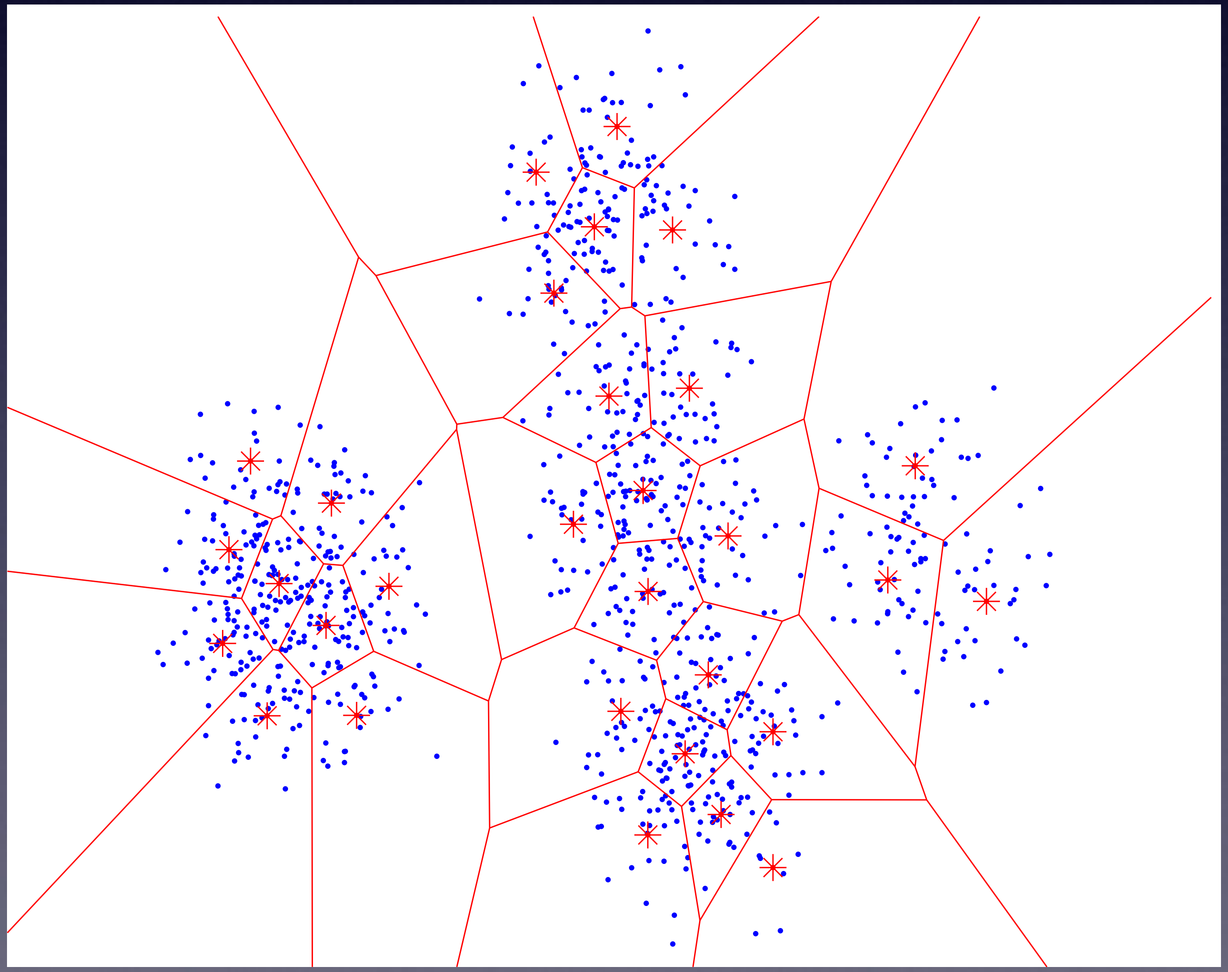
$$\hat{\mathbf{x}} = \mathbf{C}^{-1/2}(\mathbf{x} - \mu)$$

- A form of unsupervised metric learning (more on this later).

- Each descriptor is then approximated by the most similar prototype/visual word.

Visual Words

- The result of VQ is that probability of visual words is somewhat equalized (not completely).



Visual Words

- Analogy with text document matching.
- Each document (i.e. image) is represented as a vector of (TF-IDF) word frequencies (a bag of features)

$$\mathbf{v}_d = (v_1 \; \ldots \; v_K)^T \qquad\qquad v_k = \frac{N_{kd}}{N_d} \log \frac{N}{N_k}$$

- term frequency: N_{kd}/N_d (word k, document d) Nistér&Stewénius CVPR06: skip N_d.
- inverse document frequency: N/N_k - inverse frequency of word k in whole database.

Visual Words

- Image matching is done by a normalised scalar product:

$$\hat{\mathbf{v}}_q^T \hat{\mathbf{v}}_p = \cos\phi$$

- An *inverted file* makes real-time matching possible on very large datasets:

word1: frame 3, frame 17, frame 243...
word2: frame 2, frame 23, frame 33...
...

Bag of Features

- If we set TF=N_{kd}, and omit IDF we get a histogram of visual word occurrences.
- This is called a *bag-of-features/ bag-of-words/bag-of-keypoints* in the literature.
  G. Csurka et al, ”Visual Categorization with Bags of Keypoints”, ECCV’04
- The IDF weight scales each dimension separately and can be seen as a specific choice of matching metric.

Bag of Features

- The bag-of-features vector is often fed into a machine learning algorithm (LE7) or used in ANN search (LE6)

- Typically K is large and most values are zero.

  Csurka et al.’04 K=1000
  Sivic&Zisserman’03 K=6000 and 10,000
  Nistér&Stewénius’06 K=16e6

Skip interest points?

- E. Nowak, Jurie, Triggs, ”Sampling Strategies for Bag-of-Features Image Classification”, ECCV’06

- More descriptors in histogram computation result in a more informative BoF vector.

- For low-res images, number of detected points can easily be too low with standard detection thresholds.

Skip interest points?

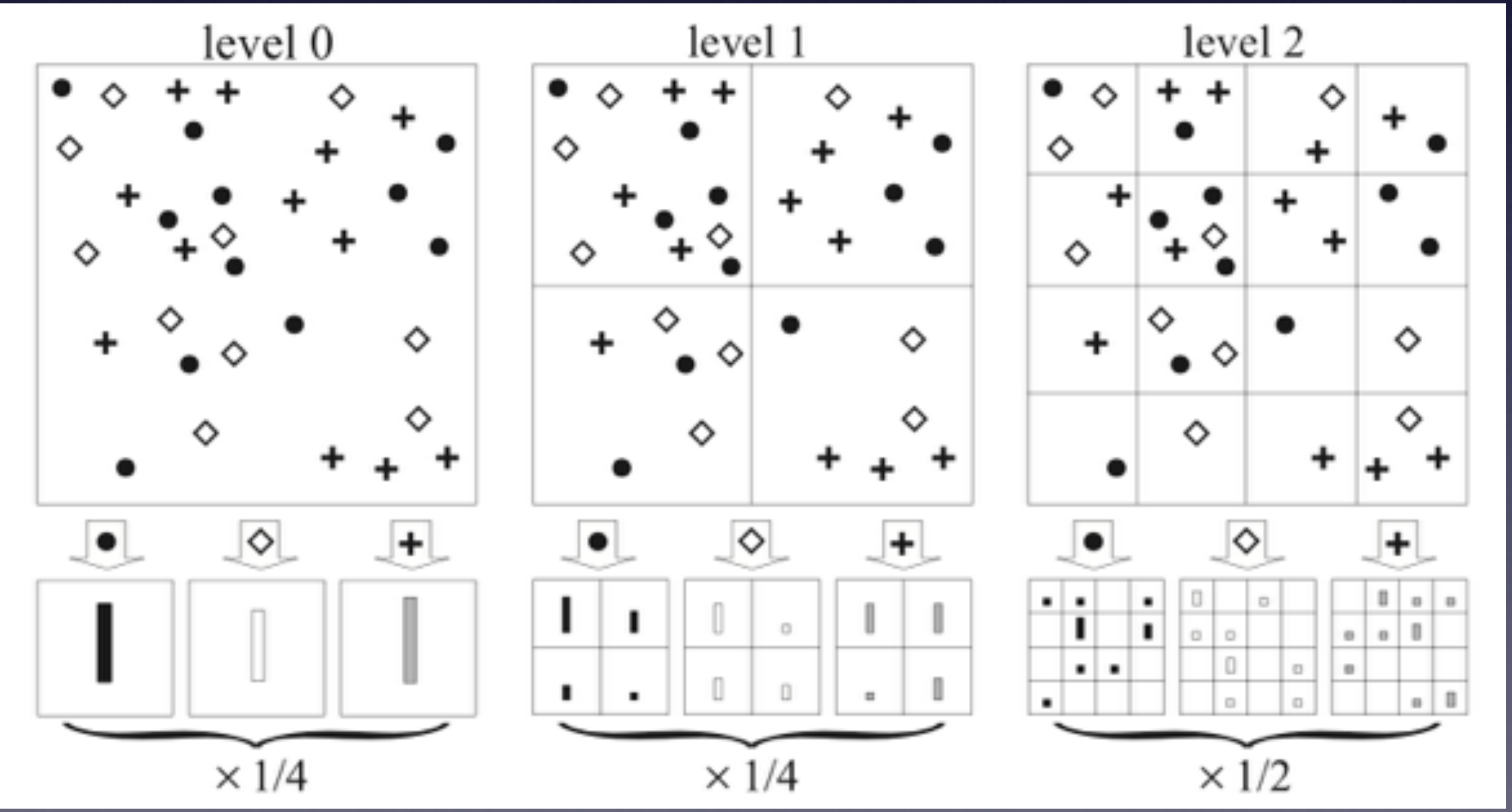
- For low detection thresholds detection is both highly biased and noisy.

Harris-Laplace | Harris-Laplace no thr | Laplace-of Gaussian | Random sampling

- Nowak, Jurie and Triggs improve performance using random sampling. Another popular choice is dense/gridded sampling.

Spatial Pyramids

- Lazebnink, Schmid &Ponce, ”Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”, CVPR’06
- Essentially: stack BoF vectors in grids of several different sizes



Spatial Pyramids

- Lazebnink, Schmid &Ponce, ”Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”, CVPR’06

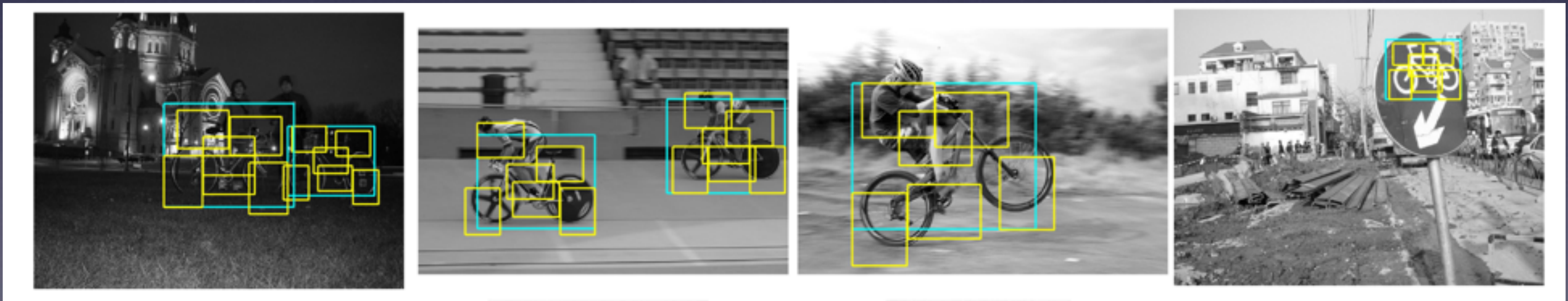
- Larger grid cells are down-weighted to compensate for the higher likelihood of matches there.

- Even with a spatial pyramid, constellation information is not fully exploited in BoF approaches, so spatial verification may be useful afterwards.

Deformable Part Models

- P. Felzenswalb et al. ”A Discriminatively Trained, Multiscale, Deformable Part Model”, CVPR’08

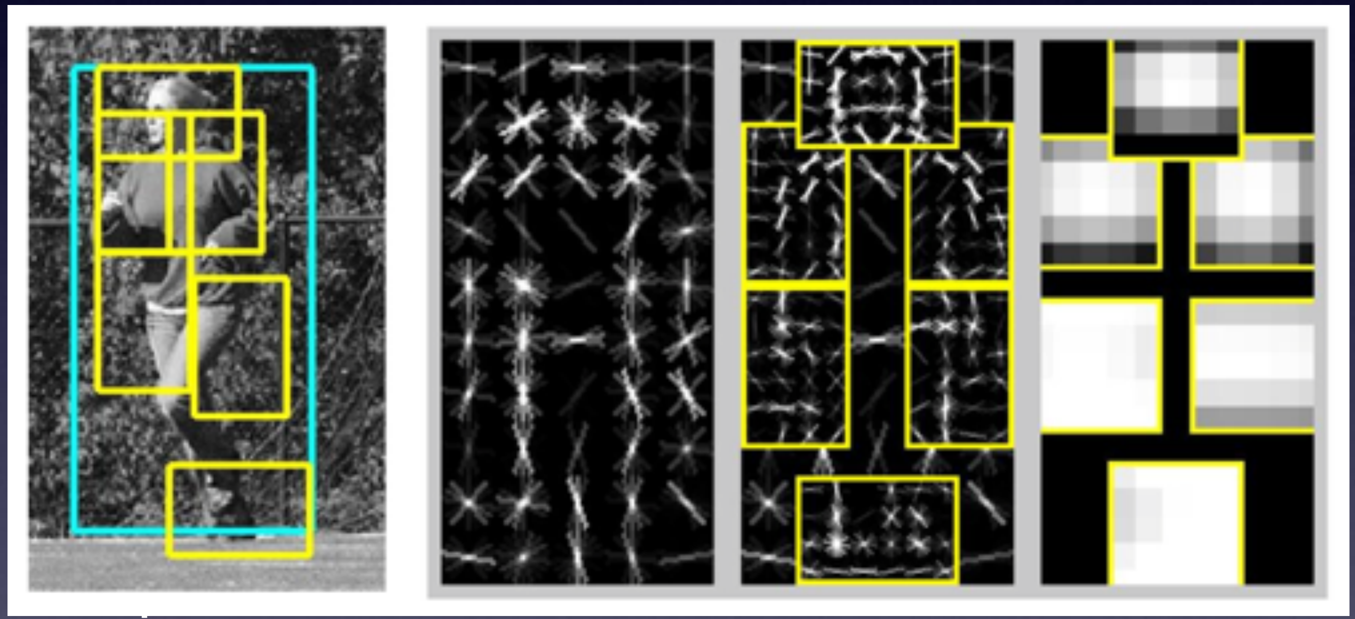
1. A coarse global model
2. A fixed number of part models with flexible spatial arrangement.



Source code available on github

Deformable Part Models

- P. Felzenswalb et al. ”A Discriminatively Trained, Multiscale, Deformable Part Model”, CVPR’08



- Detection is done on a coarse pattern
- Constellations are used as a verification - makes matching tractable.
- For several years this class of methods had the best performance in recognition contests.

Hierarchical Compositional Models

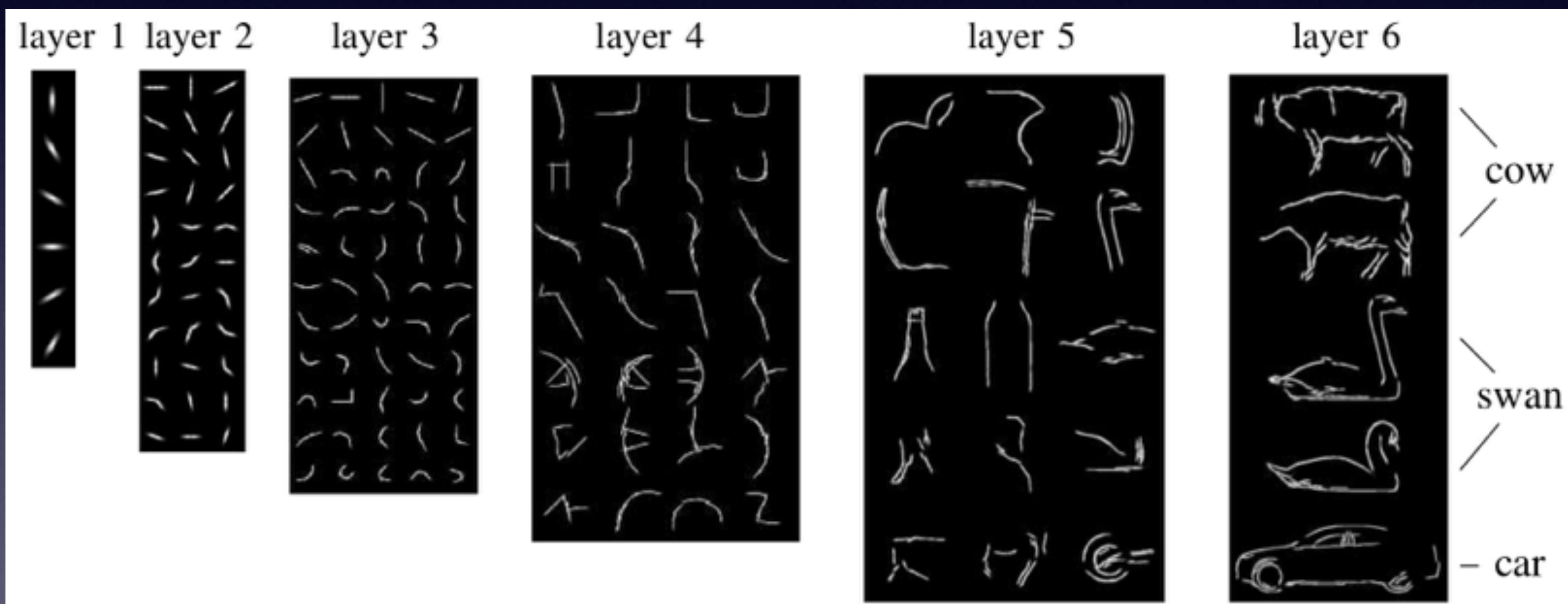
- Fidler and Leonardis, ”Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts”, CVPR’07

- Many recognition techniques (e.g. discriminative ones) are linear in the number of object categories.

- Fidler&Leonardis present an attempt at automatic feature sharing to reduce the asymptotic complexity.

Hierarchical Compositional Models

- Fidler and Leonardis, ”Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts”, CVPR’07



- Each part is a combination of parts in the previous layer. (only a subset of parts shown above for L2-L6)

Hierarchical Compositional Models

- Fidler and Leonardis, ”Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts”, CVPR’07
- Recognition is done layer by layer, by having features describe all detected L1 features in the image (a generative approach).
- Assignment in L2-L6 is done in hypothesize-verify fashion, where parts vote for constellations.
- Each constellation has flexible position and orientation of parts (amount is learned).

Hierarchical Compositional Models

- Fidler and Leonardis, ”Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts”, CVPR’07
- Learning is done incrementally, one category at a time.
- Features already present can be re-used in new categories.
- Interesting idea, but currently only contour features are used. SOTA on shape recognition 2007.

Descriptor Distances

- For a descriptor $\mathbf{q}$ in a query image. Which prototype in memory ($\mathbf{p}_1,\mathbf{p}_2,\ldots,\mathbf{p}_N$) is *most likely* to correspond to the same world object?

Descriptor Distances

- For a descriptor **q** in a query image. Which prototype in memory ($\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N$) is *most likely* to correspond to the same world object?
- Assuming additive i.i.d. Gaussian noise on all elements:

$$p(\mathbf{q}|\mathbf{p}_k) \propto \prod_{l=1}^{D} e^{-0.5(p_{kl}-q_l)^2/\sigma^2}$$

$$\max(p) \;\Leftrightarrow\; \min(-\log(p))$$

$$-\log(p(\mathbf{q}|\mathbf{p}_k) \propto \sum_{l=1}^{D}(p_{kl} - q_l)^2$$

Descriptor Distances

- So, the match with smallest distance is most likely correct, assuming i.i.d. Gaussian noise.

- What about the scalar product for normalised vectors/NCC?

$$||\mathbf{p}-\mathbf{q}||^2 = \mathbf{p}^T\mathbf{p} + \mathbf{q}^T\mathbf{q} - 2\mathbf{p}^T\mathbf{q} = 2(1-\mathbf{p}^T\mathbf{q})$$

- But are all values identically distributed?

- ...are they independent?

Chi^2 Distance

- Many descriptors (e.g. SIFT) are histogram-like in their nature.
- For histograms, the histogram values typically follow the (discrete) *Poisson distribution*: $P(k|\mu) = \mu^k e^{-\mu}/k!$
- Mean and variance:

$$E\left[P(k)\right] = \mu \qquad E\left[(P(k)-\mu)^2\right] = \mu$$

Chi² Distance

- For large values of μ, (e.g. 1000) a (continuous) Gaussian can approximate the Poisson distribution:

$$p(k|\mu) \approx \frac{1}{\mu\sqrt{2\pi}} e^{-0.5(k-\mu)^2/\mu}$$

- Again, assuming independence, this leads to a negative log likelihood proportional to:

$$-\log(p(\mathbf{q}|\mathbf{p}_k) \propto \sum_{l=1}^{D} (p_{kl} - q_l)^2/\mu_l$$

Chi^2 Distance

- If we estimate the variance by:

$$\mu_l \approx (p_{kl} + q_l)/2$$

- We find that the most likely match is the one with the smallest Chi-squared distance:

$$\mathcal{X}^2(\mathbf{q}, \mathbf{p}_k) = \sum_{l=1}^{D} \frac{(p_{kl} - q_l)^2}{p_{kl} + q_l}$$

Square root matching

- Another similar histogram measure is the square root distance:

$$d_{1/2}(\mathbf{q}, \mathbf{p}_k)^2 = \sum_{l=1}^{D} (\sqrt{p_{kl}} - \sqrt{q_l})^2$$

- Close approximation to Chi², and faster if SQRT is pre-computed (e.g. RootSIFT).

Histogram Intersection

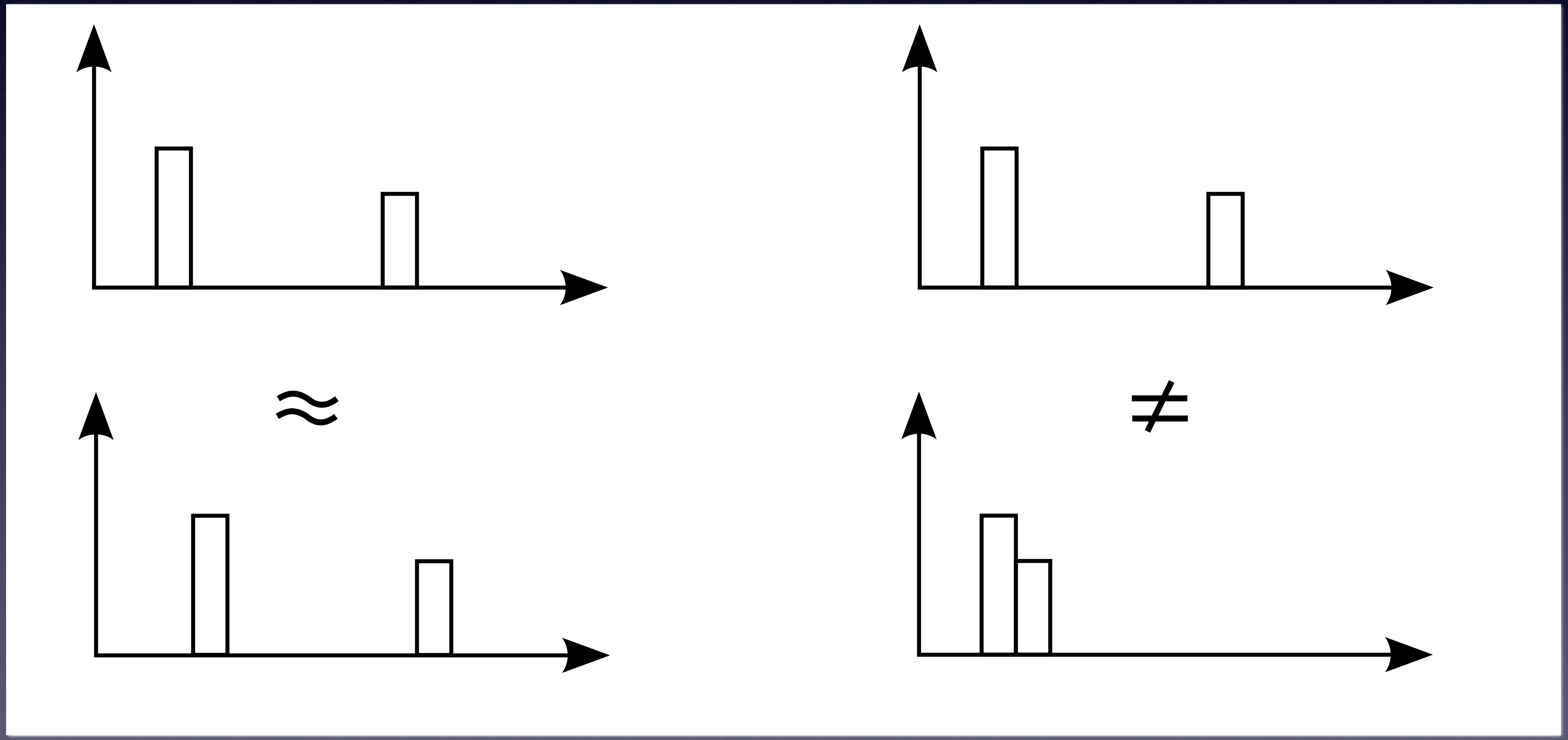
- Histogram intersection similarity measure:

$$d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{D} \min(p_i, q_i)$$

- Another common similarity measure for histogram type data.
- This far, all measures assume independence between bins.
- Good for ANN methods (LE6), but an approximation.

Earth Mover's Distance

- In histograms, neighbouring bins are typically correlated



- Instead of falling in bin i, a sample is likely to fall in bin i+1.

Earth Mover's Distance

- Distance=cost of moving values in **p** to **q**
  cost=amount*distance

- First solve a linear programming problem:
  the transportation problem, Hitchcock 1941.

$$\min_{f_{ij}} \sum_{i=1}^{D} \sum_{j=1}^{D} f_{ij} d_{ij} \quad \text{where} \quad d_{ij} = |i-j|$$

- f_{ij} amount to move from i to j.

Earth Mover's Distance

- Transportation problem, cost function:

$$\min_{f_{ij}} \sum_{i=1}^{D} \sum_{j=1}^{D} f_{ij} d_{ij} \quad \text{where} \quad d_{ij} = |i - j|$$

- Constraints:

$$f_{ij} \geq 0 \quad \forall\, i, j \in [1, D]$$

$$\sum_{i=1}^{D} f_{ij} = q_j \quad \forall j \in [1, D]$$

$$\sum_{i=1}^{D} f_{ij} \leq p_j \quad \forall j \in [1, D]$$

Earth Mover's Distance

- Now compute EMD as:

$$d(\mathbf{p}, \mathbf{q}) = \min_{f_{ij}} \frac{\sum_{i=1}^{D} \sum_{j=1}^{D} f_{ij} |i - j|}{\sum_{i=1}^{D} \sum_{j=1}^{D} f_{ij}}$$

- The denominator is needed if histograms are computed from variable numbers of samples.

- Inroduced in Computer Vision by:
Y. Rubner, C. Tomasi, and L. J. Guibas. "The earth mover's distance as a metric for image retrieval". IJCV Nov 2000

- Local expert: Thomas Kaijser

Pyramid Match Kernel

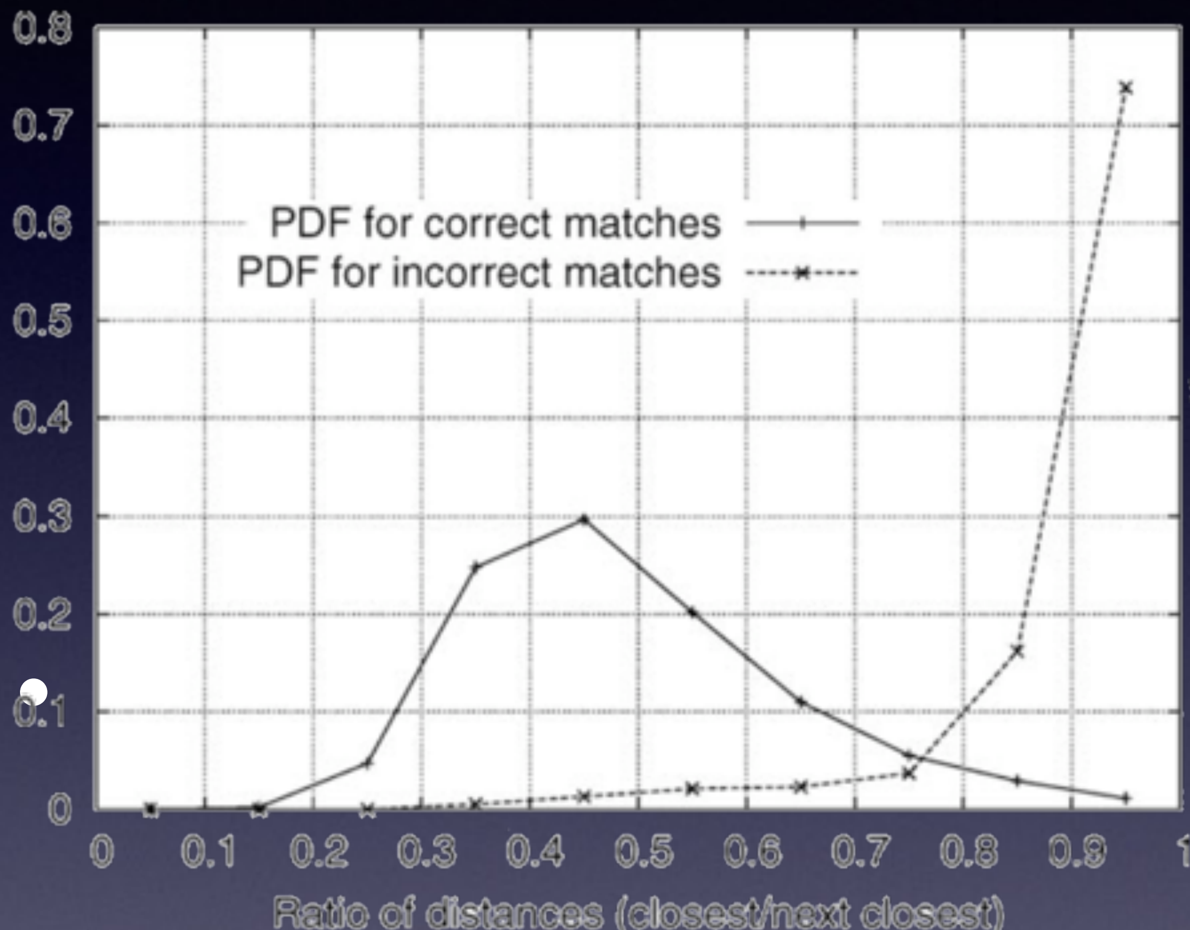
- EMD approximation:
  Grauman&Darrell, ICCV'05, "Pyramid Match Kernels: Discriminative Classification with sets of image features", ICCV05
- Create "scale pyramid" where bins are hierarchically grouped.
- Downweight coarser scales in a way that ensures Mercer kernel properties (needed for SVM convergence).
- Spatial pyramid for BoF was formulated using PMK.

Ratio Score

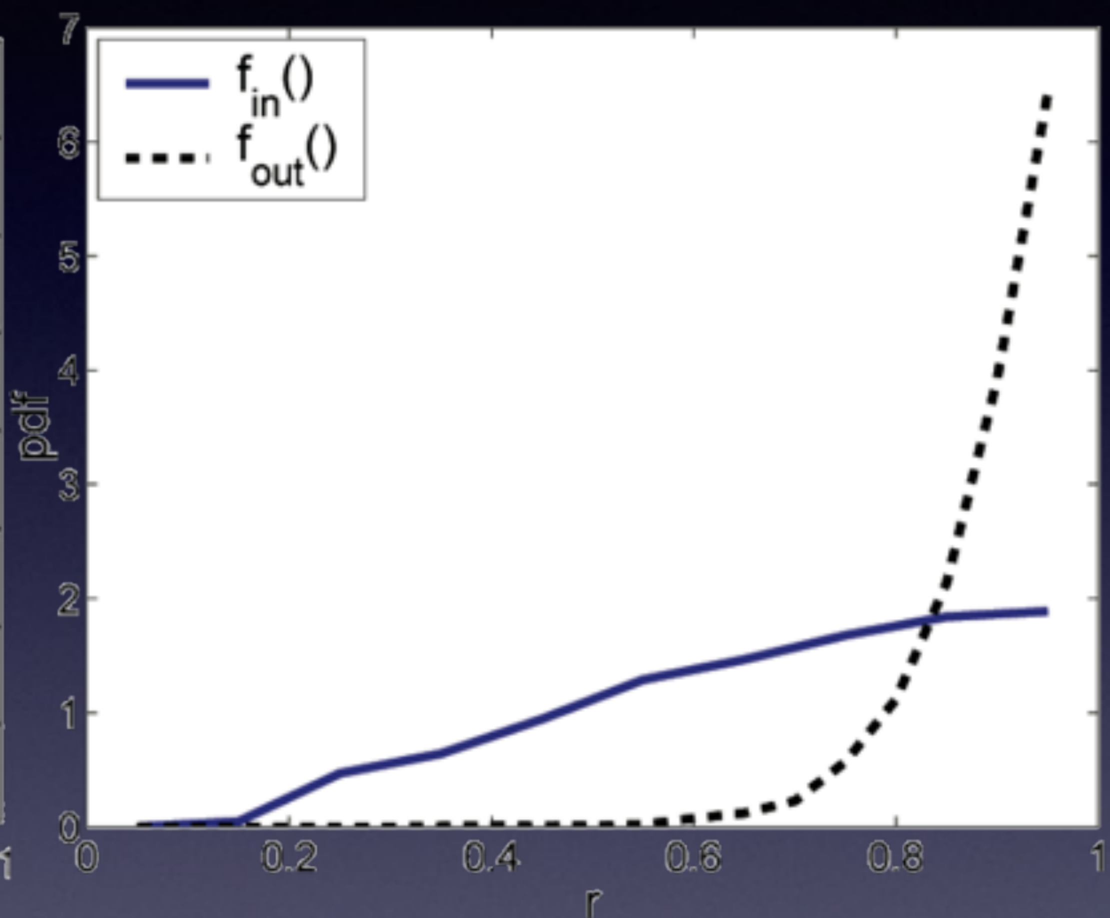
- If we have best matches for descriptors $\mathbf{q}_1$ and $\mathbf{q}_2$ in the image. Which one is better?
- Both **similarity** and **risk of misclassification** matter!
- Scoring the match for $\mathbf{q}_1$, according to the ratio between the best, and the second best match compensates for this risk:

$$r = d_1/d_2$$

Ratio Score



Lowe IJCV04



Goshen&Shimshoni PAMI08

Learning the Metric

- What we ultimately want is to distinguish good feature matches from bad.

- Collect known corresponding descriptors:

$$\{(\mathbf{p}_k, \mathbf{q}_k)\}_1^K \quad \text{and set} \quad \mathbf{d}_k = \mathbf{p}_k - \mathbf{q}_k$$

- We now want to find a linear transformation that makes the noise equal in magnitude in all directions:

$$\mathbf{y}_k = \mathbf{T}\mathbf{p}_k \quad \text{assuming} \quad \mathbf{d}_k \sim \mathcal{N}(0, \mathbf{C})$$

Learning the Metric

- Find a whitening transform **T** from the covariance matrix:

$$\mathbf{C} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{d}_k\mathbf{d}_k^T \quad \text{with} \quad \mathbf{TCT}^T = \mathbf{I}$$

- Valid solutions:

$$\mathbf{T} = \mathbf{RC}^{-1/2} \quad \text{where} \quad \mathbf{RR}^T = \mathbf{I}$$

- If we only use the first few dimensions we should choose **R** such that it selects dimensions where we “see things happen”.

Learning the Metric

- Find **R** from PCA of transformed SIFT feature space:

$$\mathbf{C}_b = \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T - \mathbf{m}\mathbf{m}^T \qquad \mathbf{m} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n$$

$$\mathbf{R}\mathbf{D}\mathbf{R}^T = \mathbf{C}$$

- Final contraction operator:

$$\mathbf{P} = \bar{\mathbf{I}}_k\mathbf{R}\mathbf{C}^{-1/2}$$

- Where $\mathbf{I_k}$ is a k*128 truncated identity matrix.

Learning the Metric

- This Mahalanobis metric for features was published at ICCV07 by Mikolajczyk&Matas, SIFT 128→40 dim
- A similar method that only finds a rotation called linear discriminant embedding(LDE) also at ICCV07 by Hua&Brown&Winder,
SIFT128→14/18dim
- Besides reducing dimensionality, these techniques also improve matching results.

Learning the Metric

- Linear Discriminant Embedding(LDE)
- Maximise

$$\mathbf{J}(\mathbf{w}) = \frac{\sum_{\text{outlier}(i,j)} \mathbf{w}^T(\mathbf{p}_i - \mathbf{q}_j)^2}{\sum_{\text{inlier}(i,j)} \mathbf{w}^T(\mathbf{p}_i - \mathbf{q}_j)^2}$$

$$\mathbf{J}(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{A}\mathbf{w}}{\mathbf{w}^T\mathbf{B}\mathbf{w}}, \quad ||\mathbf{w}|| = 1$$

- Where **A** covariance for outliers and **B** inliers.

Learning the Metric

- J(**w**) is maximised by eigenvectors with large eigenvalues in $\mathbf{B}^{-1}\mathbf{A}$
- Eigenvalues of **B** are set to $\tilde{\lambda}_i = \max(\lambda_i, \lambda_r)$

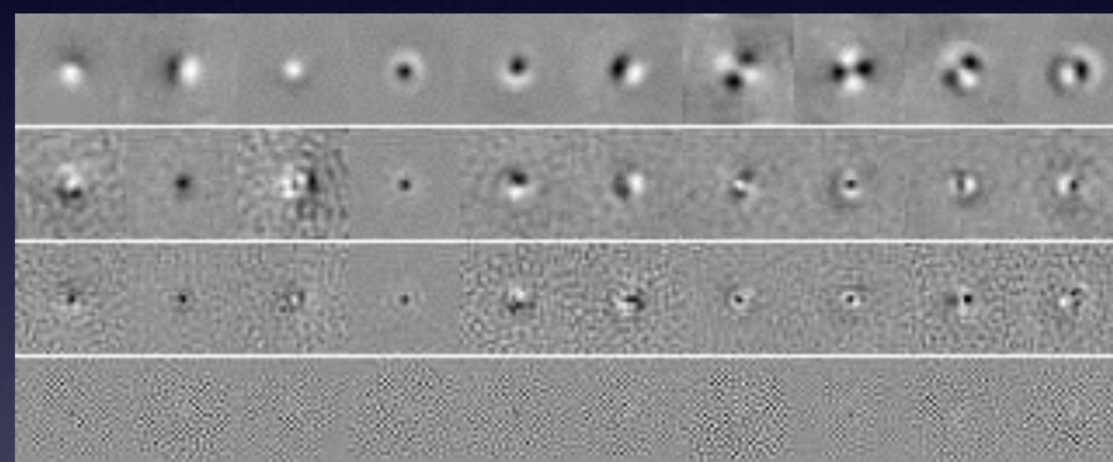
$$r = \arg\min_n \frac{\sum_{i=1}^{N} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \geq \alpha$$

- $\boldsymbol{\alpha}$ can be interpreted as a threshold on SNR. This is called *Power Regularisation*
- Many variations of the algorithm in the paper.

Learning the Metric

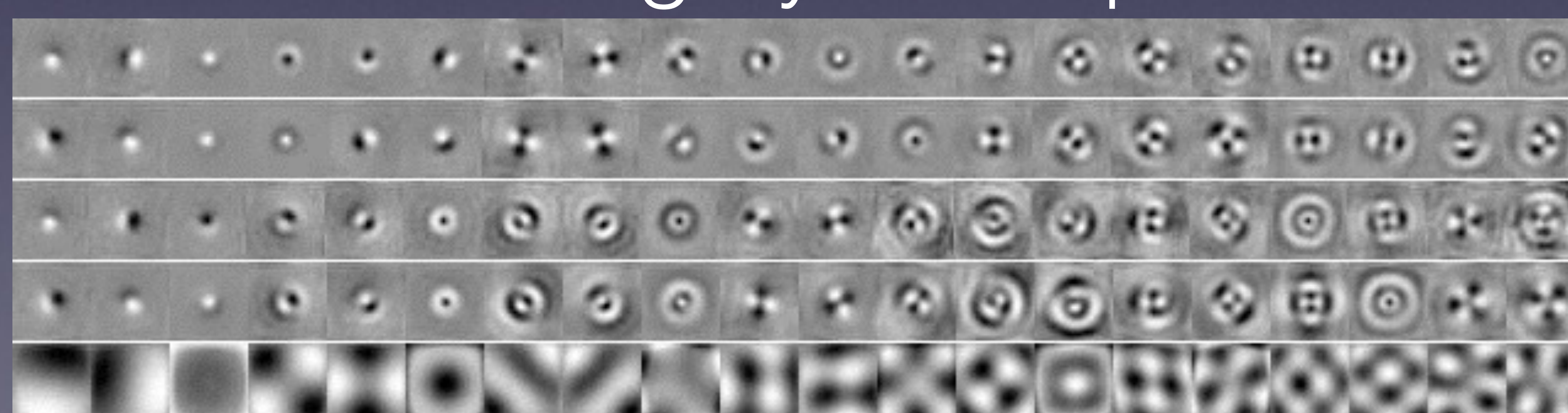
- Some LDE results on grey-scale patches:
  Reducing the amount of power reg:

20%
10%
2%
0%



- Linear filters found on grey-scale patches:

LDE-I
LDE-II
OLDE-I
OLDE-II
PCA



Discussion

- Questions/comments on today's paper:

  J. Sivic, A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos", ICCV 2003

Paper for next week

- Paper to read for next week:

  M. Muja and D.G. Lowe, ”Scalable Nearest Neighbour Algorithms for High Dimensional Data”, TPAMI 2014

- NB! Journal paper, so longer than previous papers.